Supernova Recognition using Support Vector Machines

Raquel Romano
Computational Research Division
Lawrence Berkeley National Laboratory
romano@hpcrd.lbl.gov
http://vis.lbl.gov/~romano

September 20, 2006

Type Ia Supernovae

- Stellar explosions appearing as bright spots near galaxies
- Rare: 1-2 per millenium
- Random and fleeting: wax and wane within several weeks

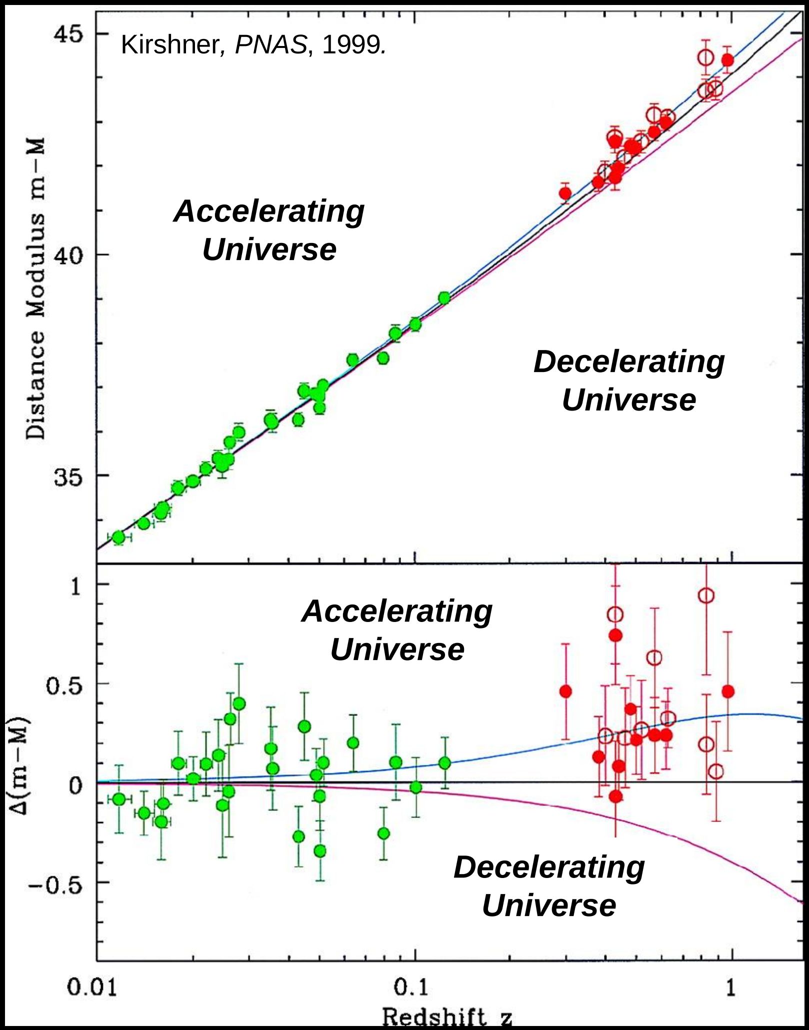




BERKELEY LAB

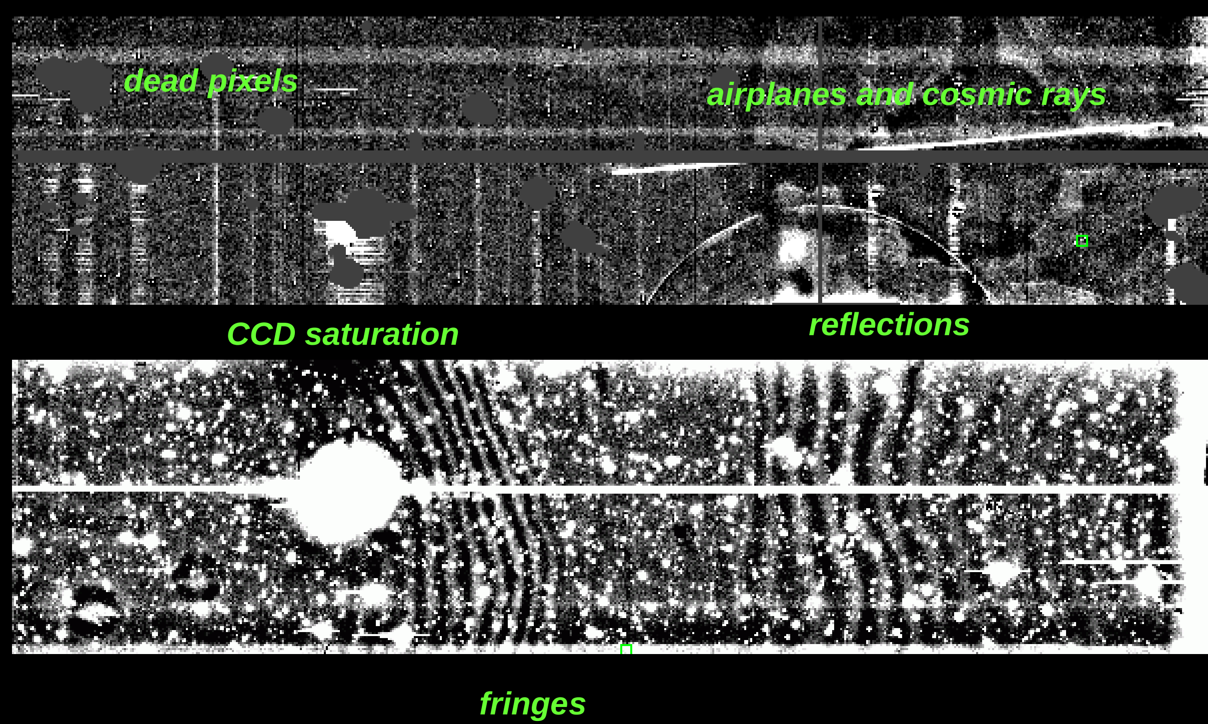
Type Ia Supernovae Studies

- First direct experimental evidence that universe is accelerating
- Uniform peak brightness, optical spectra, and light curves
- Time-varying spectra of thousands of Type Ia supernovae needed to constrain estimate of acceleration rate



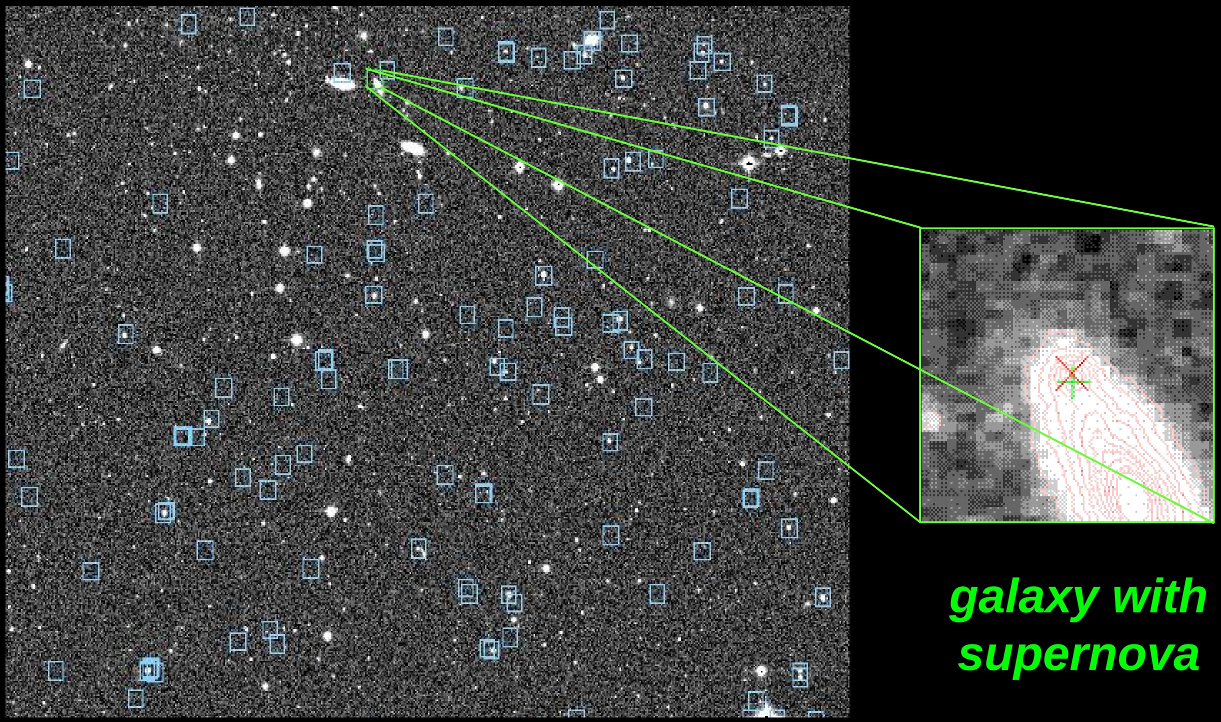
Computational Difficulty

- Noisy imagery with many artifacts
- Large data sets captured and analyzed nightly: ~30,000 images/night (85 Gb) and growing



Computational Difficulty

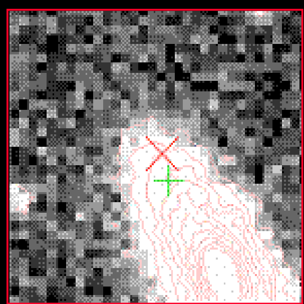
- Noisy imagery
- Large data sets
- Important to avoid missing a single candidate
- False detections make human workload burdensome
- Early detection is critical, but difficult due to faintness

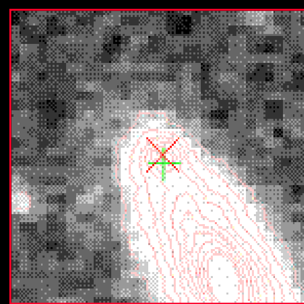


galaxy with supernova

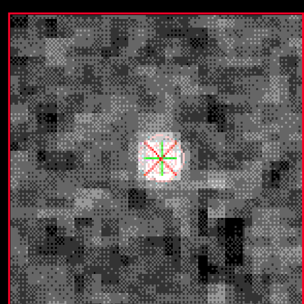
Supernova Detection in Astronomical Imagery

reference ***new***





subtraction



$$\begin{bmatrix} 11.88 \\ 7.847 \\ -0.117 \\ 3.269 \\ 3.22 \\ 10.36 \\ 9.692 \\ -0.1523 \\ 0.7023 \\ \vdots \\ 0.008165 \\ 0.05337 \end{bmatrix}$$

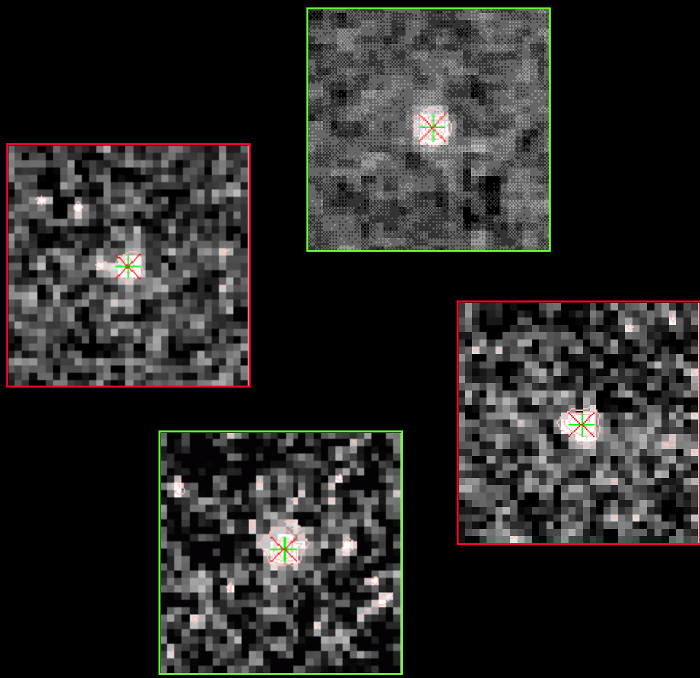
- Reference image subtracted from new observed image
- Geometric and photometric features computed from subtraction subimage
- Manually tuned upper/lower thresholds on features determine candidates
- Final decision made by human scanners (postdocs, senior scientists)

The Nearby Supernova Factory
http://snfactory.lbl.gov

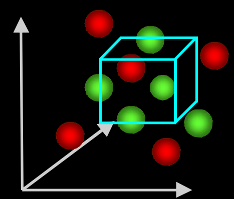
Thresholding is Fragile

Frequent adjustments to thresholds to simultaneously minimize

- False Detections (variable stars, asteroids, image artifacts, *i.e.,* junk)
- Missed Supernovae

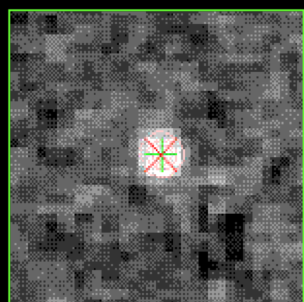


$$\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix}$$

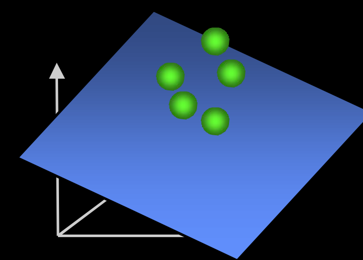


Supervised Learning Problem

- Features computed from each candidate subimage: mapping to *n*-dimensional space



$$\longrightarrow \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix} \longrightarrow$$

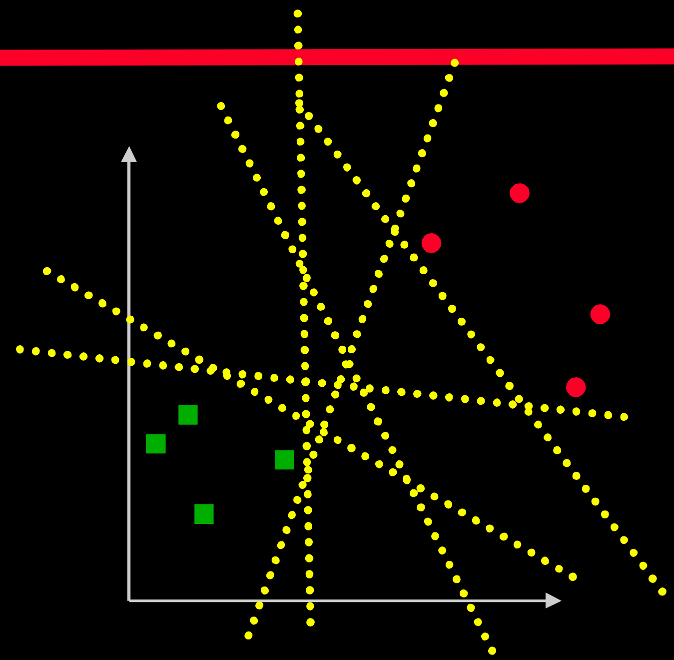


- Human scanning provides labels for positive and negative examples
- Compute “optimal” decision boundary in feature space
- Considerations:
  - Complexity of decision boundary
  - Separability of classes in feature space

Linear SVM: Margin Maximization

Linear SVM

- compute the optimal separating hyperplane between data points belonging to two classes

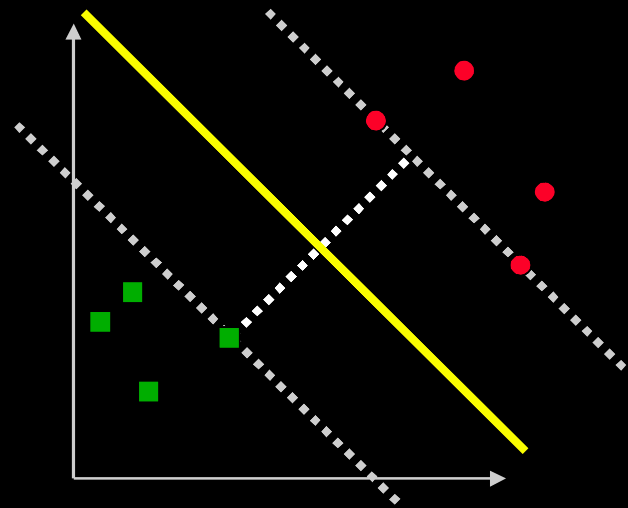


Chen, Lin, Schölkopf, A Tutorial on ν–Support Vector Machines, 2003.

Linear SVM: Margin Maximization

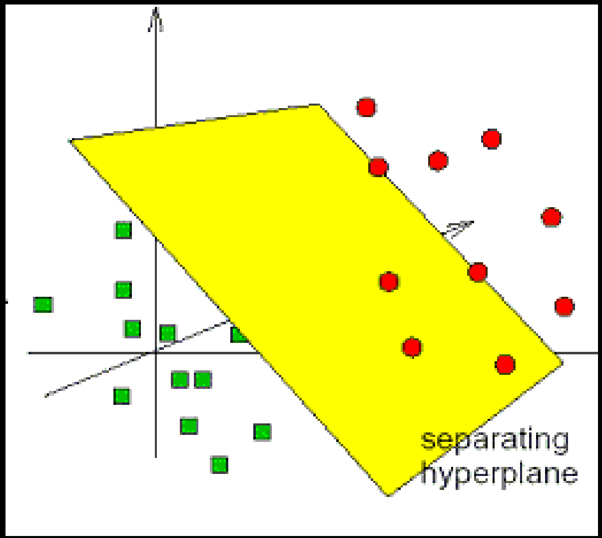
Linear SVM

- compute the optimal separating hyperplane between data points belonging to two classes



Optimal Hyperplane

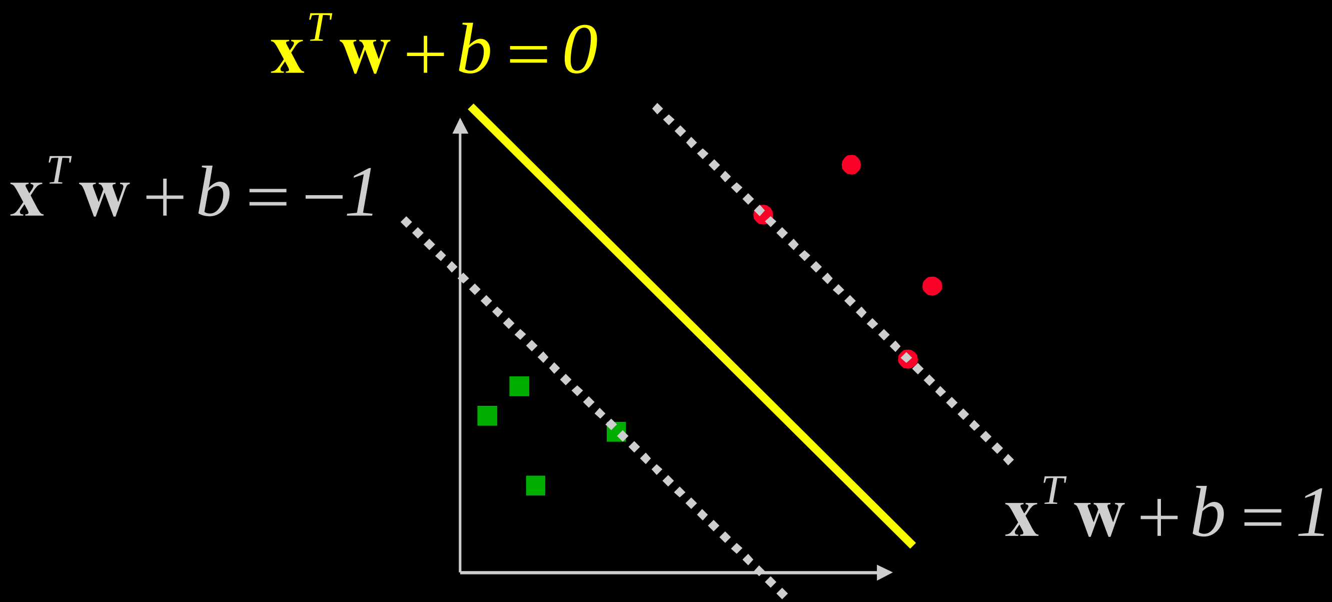
- maximize the distance from the decision surface to the nearest point in each class
- orthogonal to shortest line between convex hulls
- maximum margin separation



BERKELEY LAB

Chen, Lin, Schölkopf, A Tutorial on ν–Support Vector Machines, 2003.

Margin Maximization



- Maximize distance between two parallel supporting planes
- Distance = margin = $\dfrac{2}{\|\mathbf{w}\|}$

Chen, Lin, Schölkopf, A Tutorial on ν–Support Vector Machines, 2003.

Constrained Optimization Problem

Objective Function

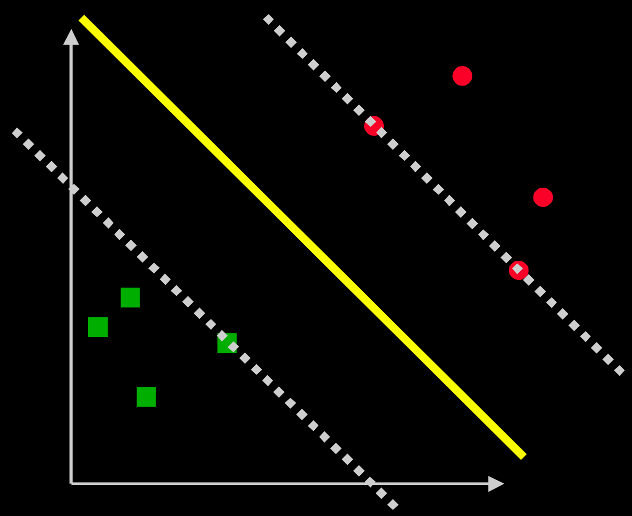
$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to}$$

$$\mathbf{x}_i^T\mathbf{w} + b \geq 1 \quad \text{for } i \in c_+$$

$$\mathbf{x}_i^T\mathbf{w} + b \leq -1 \quad \text{for } i \in c_-$$

Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i (y_i(\mathbf{x}_i^T\mathbf{w} + b) - 1)$$



- $y_i = 1$ for $i \in c_+$
- $y_i = -1$ for $i \in c_-$

Chen, Lin, Schölkopf, A Tutorial on ν–Support Vector Machines, 2003.

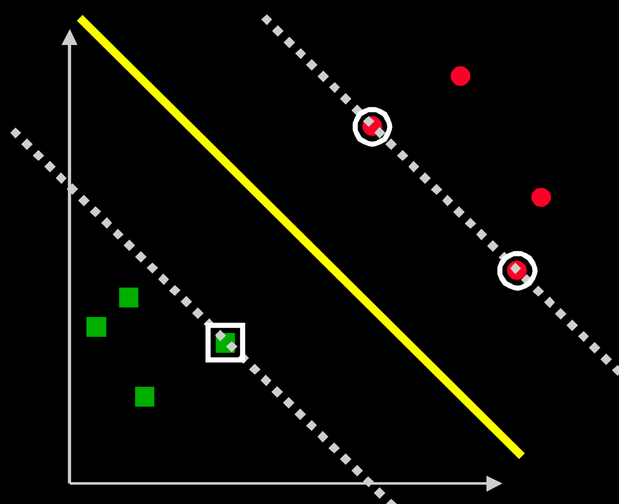
Lagrangian Formulation

Lagrangian $$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i (y_i(\mathbf{x}_i^T \mathbf{w} + b) - 1)$$

Differentiate $$\frac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0, \quad \frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0$$

$$\Rightarrow \sum_i \alpha_i y_i = 0, \quad \mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

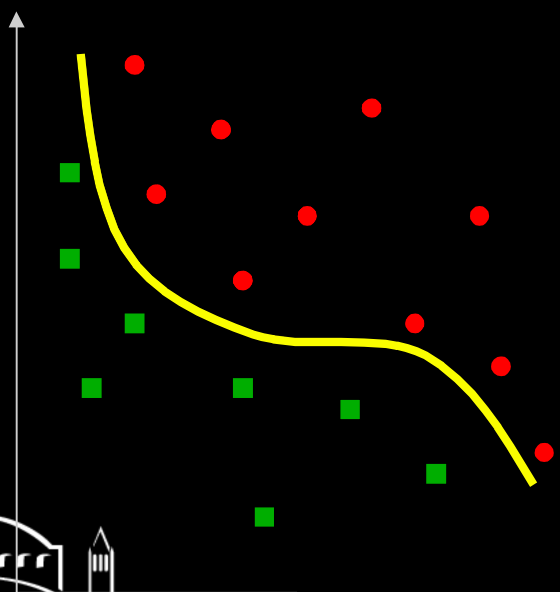
Solution depends only on ***support vectors*** $\mathbf{x}_i$ s.t. $\alpha_i > 0$



Chen, Lin, Schölkopf, A Tutorial on ν–Support Vector Machines, 2003.

Nonlinear Decision Boundary

- Reality: classes may not be linearly separable



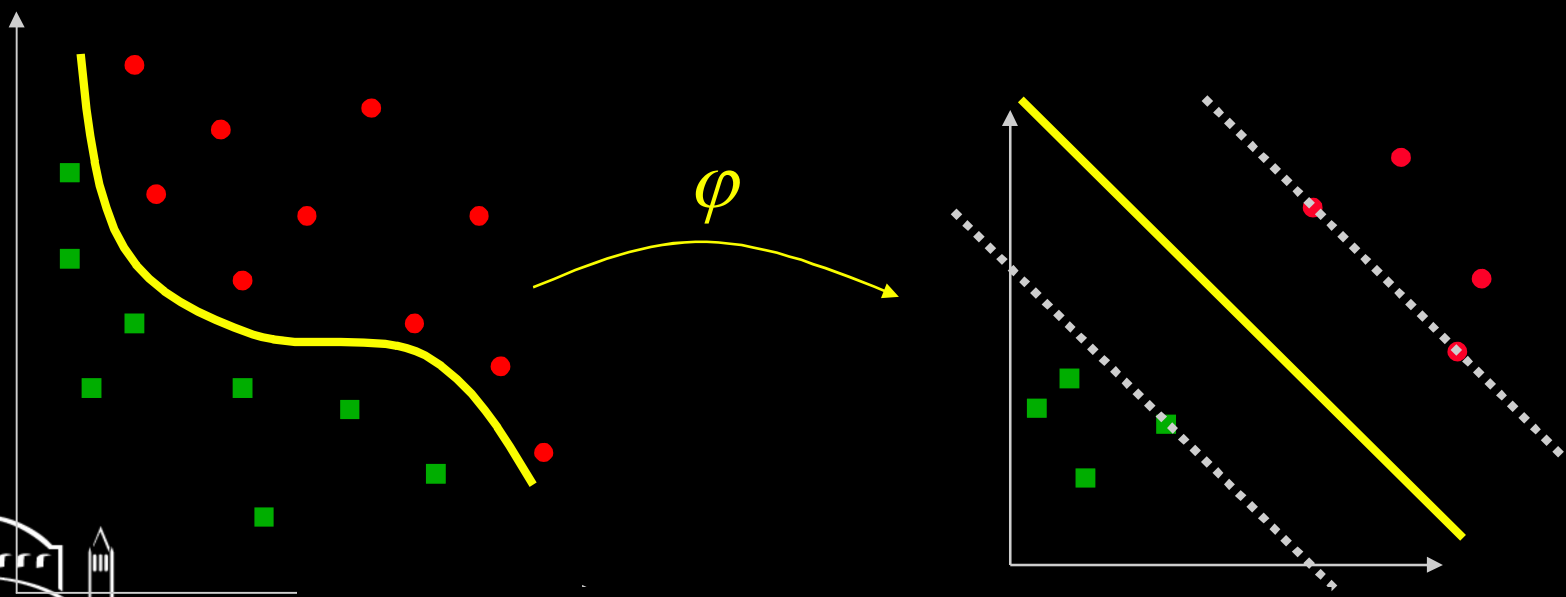
Chen, Lin, Schölkopf, A Tutorial on ν–Support Vector Machines, 2003.

Nonlinear Decision Boundary

- Reality: classes may not be linearly separable
- Map points to a higher-dimensional feature space

  e.g. products up to degree d

$$\mathbf{x}^T = [x_1, x_2] \qquad \varphi(\mathbf{x})^T = [x_1{}^2, x_1 x_2, x_2{}^2]$$



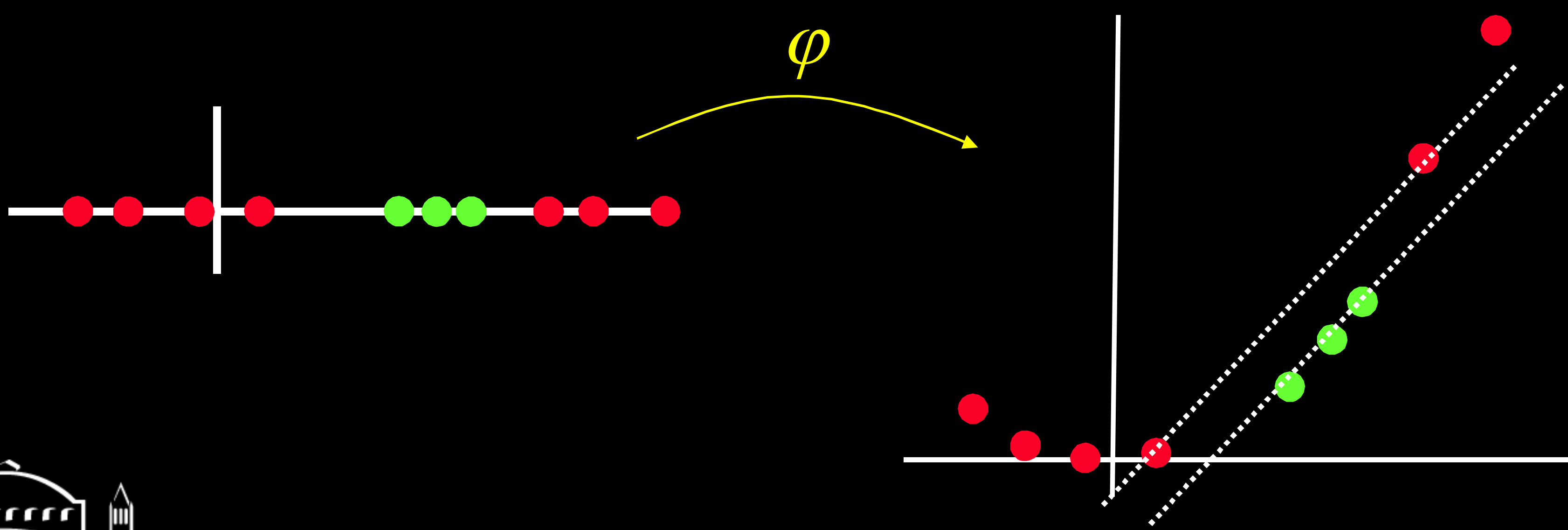
Chen, Lin, Schölkopf, A Tutorial on ν–Support Vector Machines, 2003.

Nonlinear Decision Boundary

- Map points to a higher-dimensional feature space
- Generalized inner product is called a ***kernel***

$$\mathbf{x}^T = [x_1, x_2] \qquad \varphi(\mathbf{x})^T = \left[x_1^{\;2}, x_1 x_2, x_2^{\;2}\right]$$

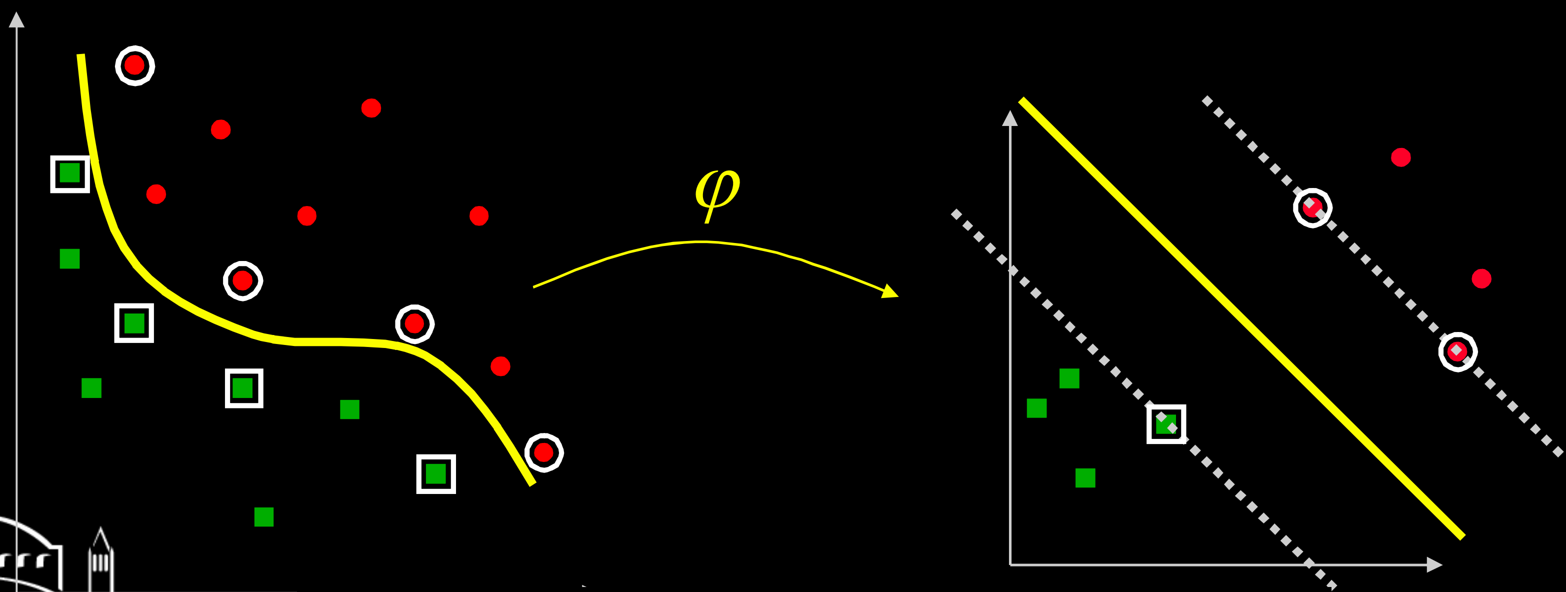
$$\mathbf{x}^T \mathbf{w} = x_1 w_1 + x_2 w_2 \qquad \textbf{\textit{kernel}} \quad k(\mathbf{x}, \mathbf{w}) = \varphi(\mathbf{x})^T \varphi(\mathbf{w})$$



Chen, Lin, Schölkopf, A Tutorial on ν–Support Vector Machines, 2003.

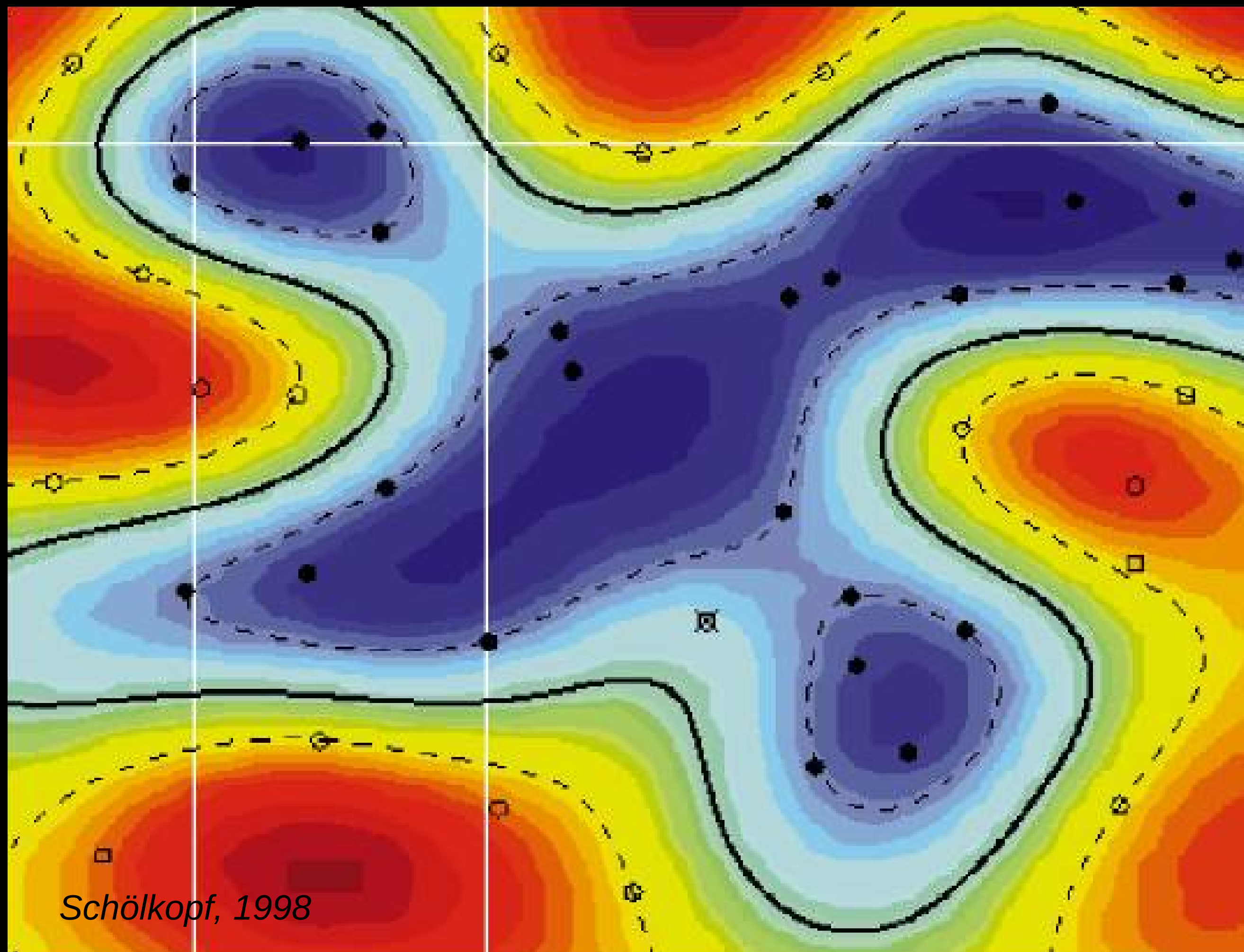
Kernels

- Can explicitly define kernel $\varphi(\mathbf{x})^T \varphi(\mathbf{w}) = k(\mathbf{x}, \mathbf{w})$ to induce implicit mapping φ
- Gaussian radial basis function $k(\mathbf{x}, \mathbf{w}) = \exp(-\frac{\| \mathbf{x} - \mathbf{w} \|^2}{2\sigma^2})$
- Decision boundary is a linear combination of support vectors, optimally chosen from training set



Chen, Lin, Schölkopf, A Tutorial on ν–Support Vector Machines, 2003.

Example: 2D RBF

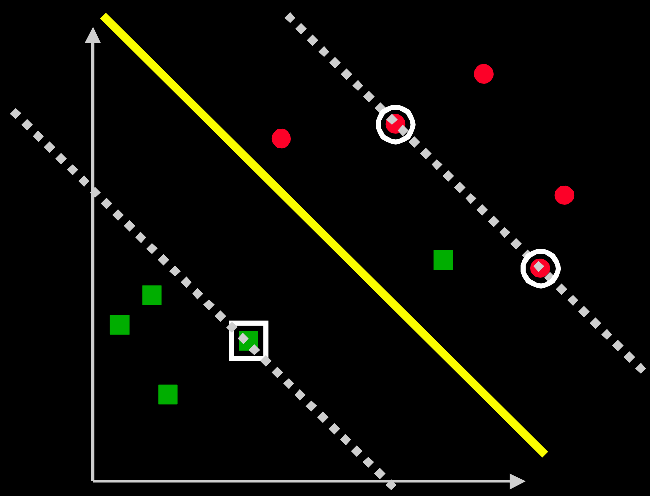


Handling Real Data

- No separating hyperplane, even after mapping!
- **Soft margin classifiers**
  - Slack variables allowing points to lie inside margin:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i} \xi_i \ \text{ subject to } \ y_i(\mathbf{x}_i^T\mathbf{w}+b) \geq 1 - \xi_i$$

  - Or: penalty terms for number of examples that are support vectors, number of examples on wrong side of hyperplane



Application to Supernova Recognition

- Overfitting:

  trade-off between complexity of boundary and generalization error

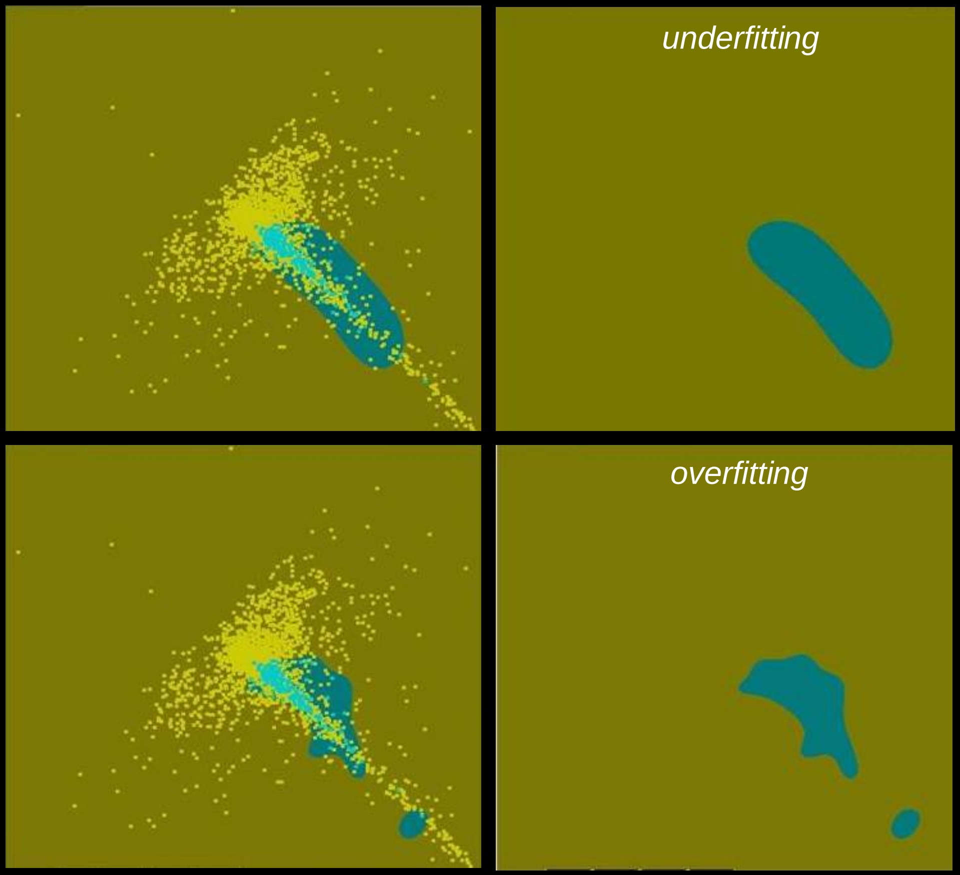
- Parameter-selection:
  - variance on Gaussian kernels
  - constants on soft margin terms in objective function

+ examples (accepted candidates)

- examples (rejected candidates)

+ region

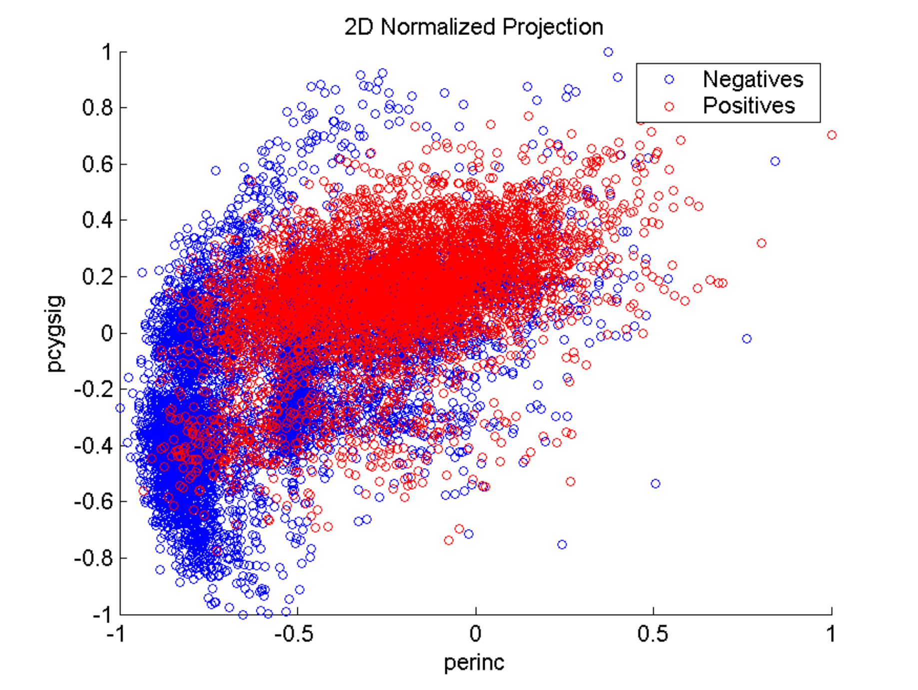
- region

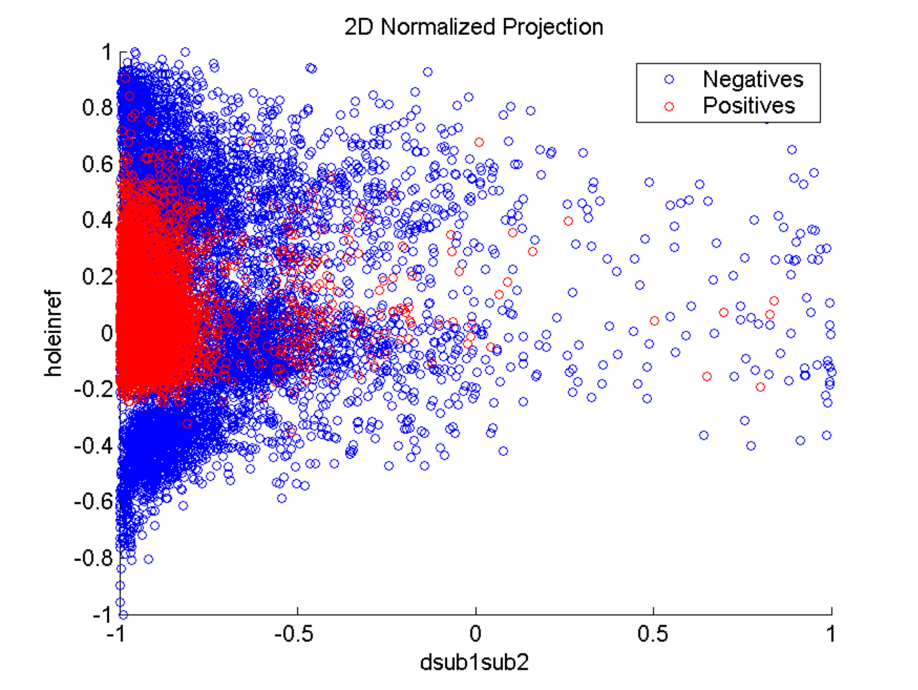


LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm

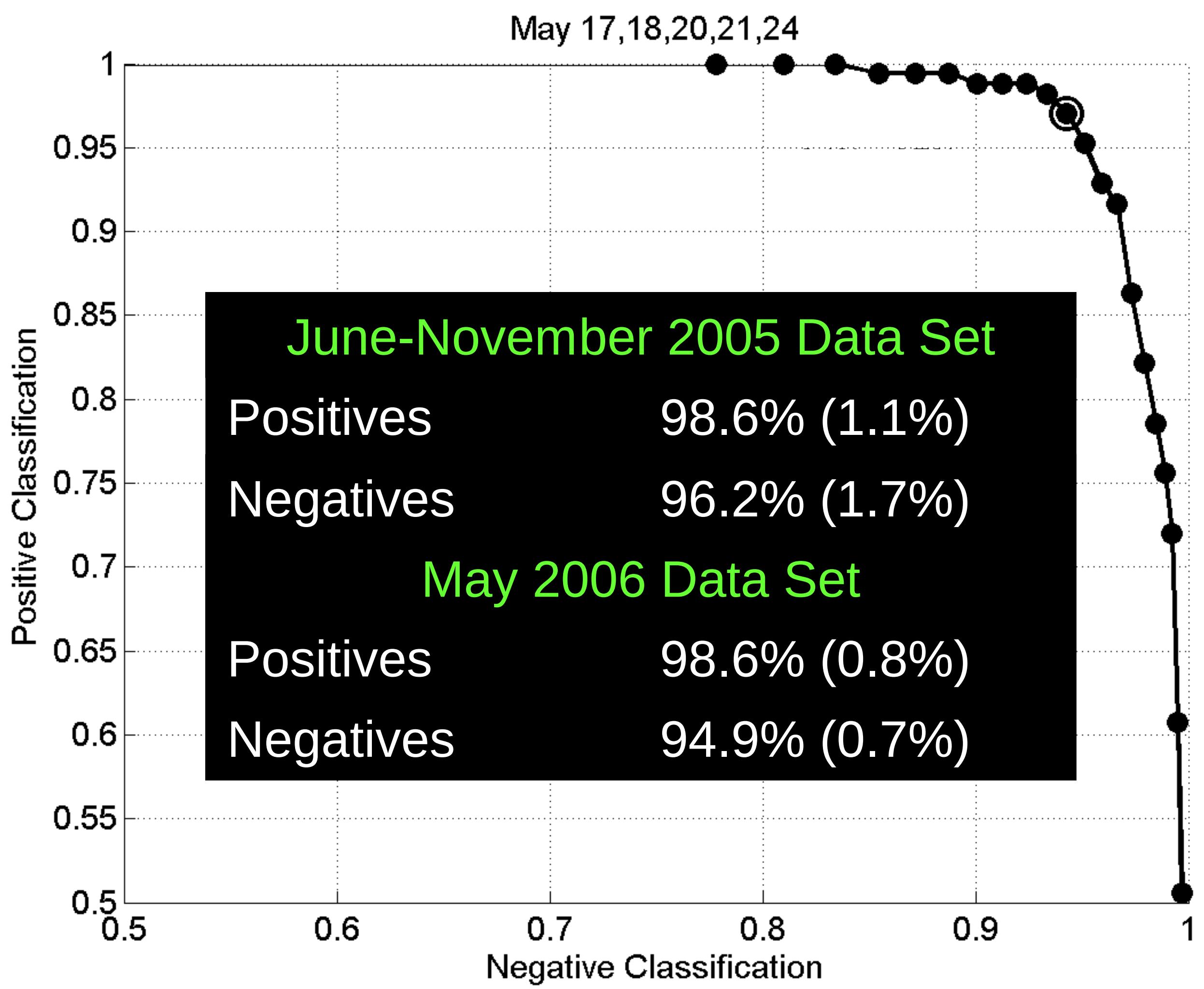
Imbalanced Data & Class Uncertainty

- Ratio of positives to negatives less than 1/10,000
- Many negatives in region of overlap
- Potentially mislabeled examples





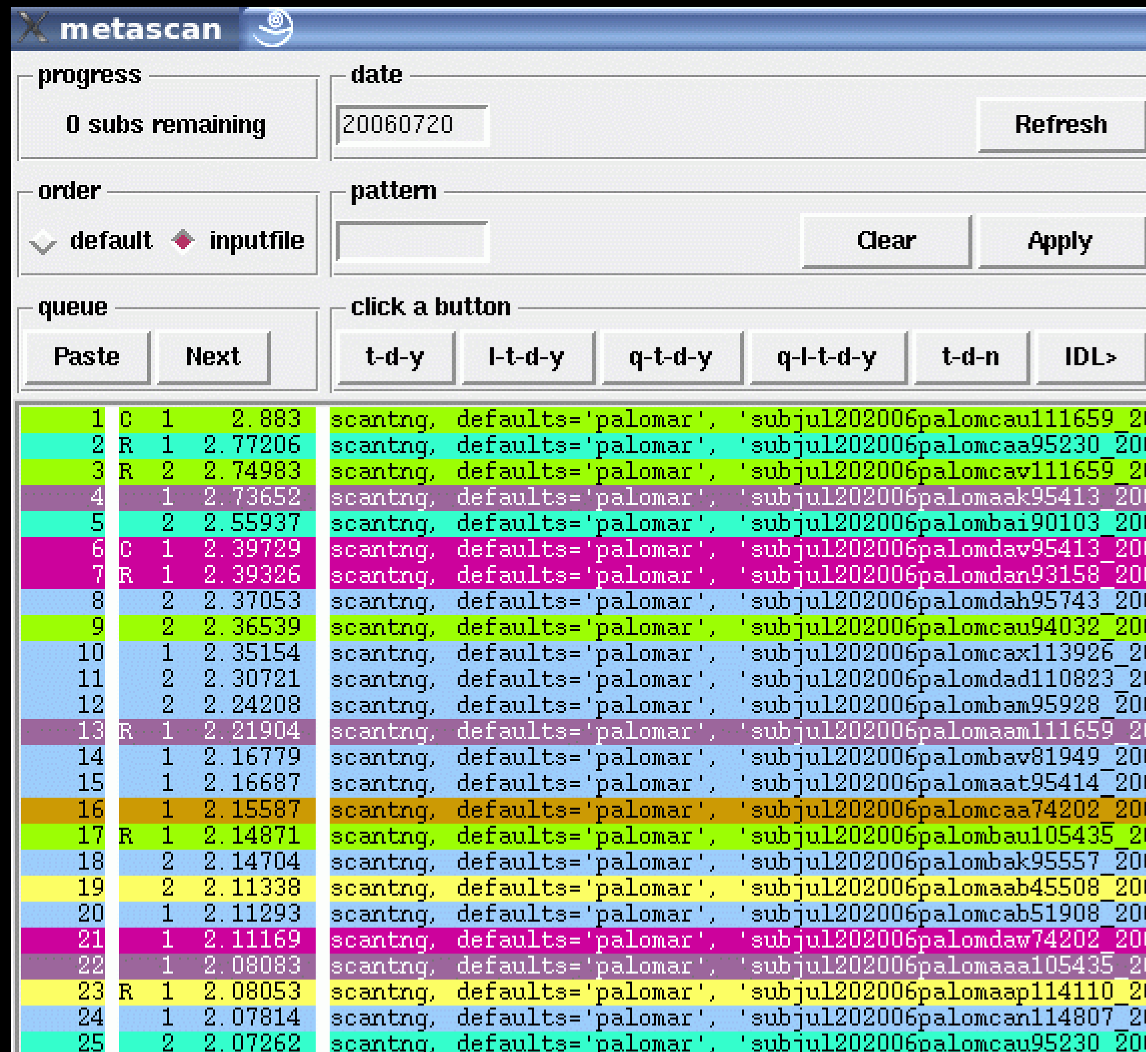
5-Fold Cross-Validation Tests



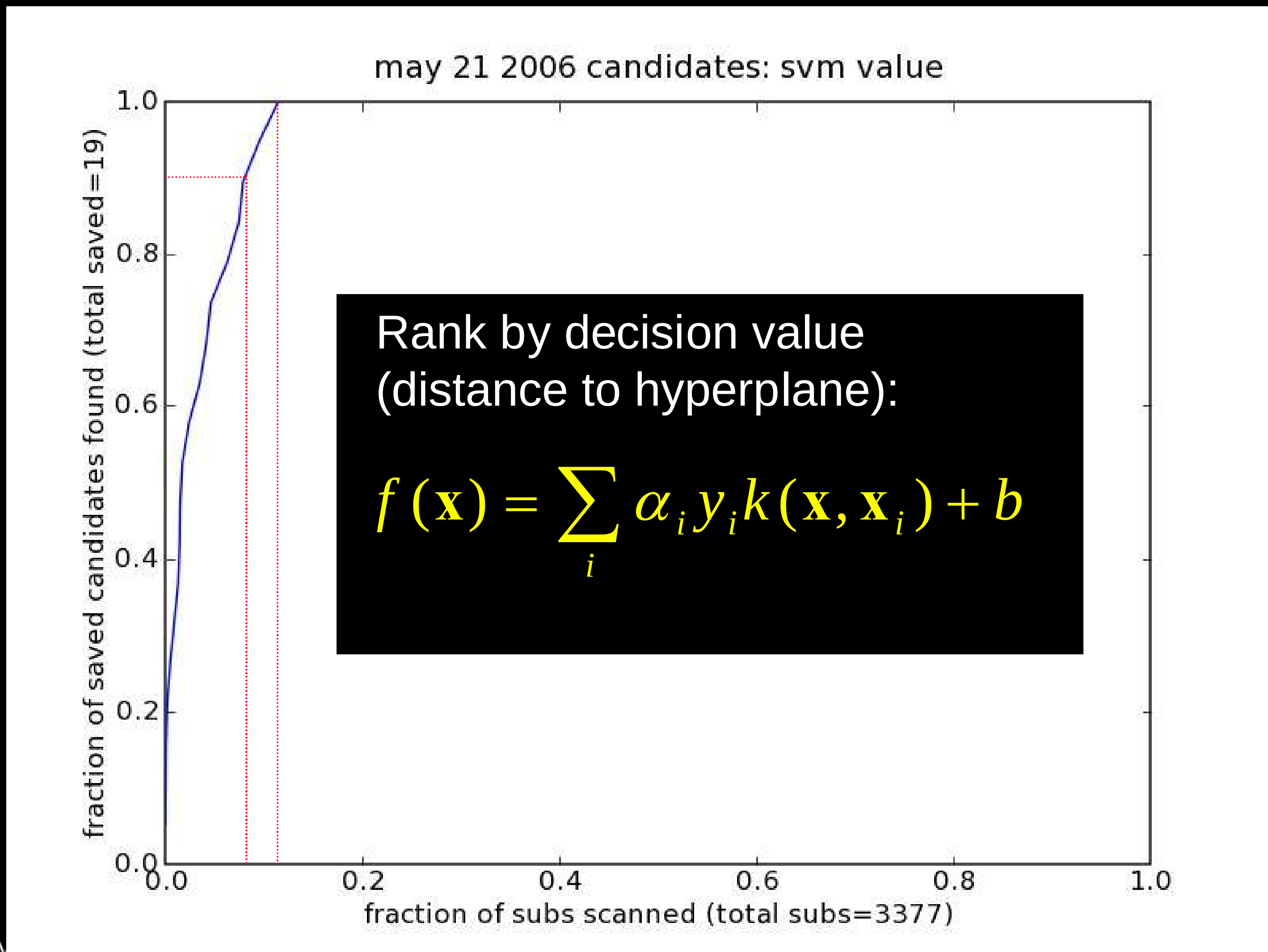
June-November 2005 Data Set	
Positives	98.6% (1.1%)
Negatives	96.2% (1.7%)
May 2006 Data Set	
Positives	98.6% (0.8%)
Negatives	94.9% (0.7%)

Incremental Sampling

- ~500-1000 subs to scan each morning: sorted by SVM decision value
- Time-sensitive: find good candidates early to schedule follow-up observations that night
- Labels from high-ranking examples: refine decision boundary



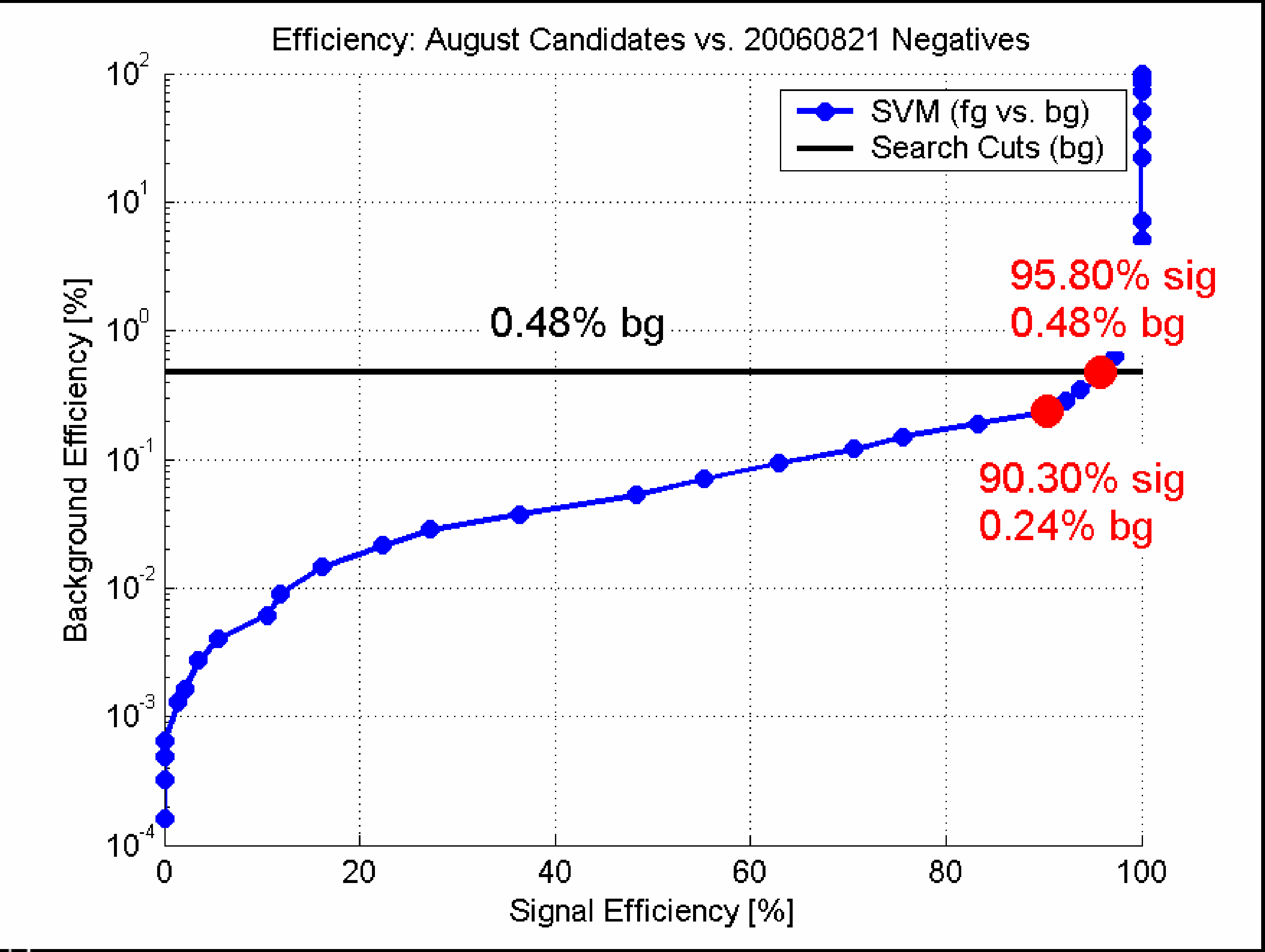
Impact on Supernova Search



Rank by decision value
(distance to hyperplane):

$$f(\mathbf{x}) = \sum_i \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b$$

Impact on Supernova Search

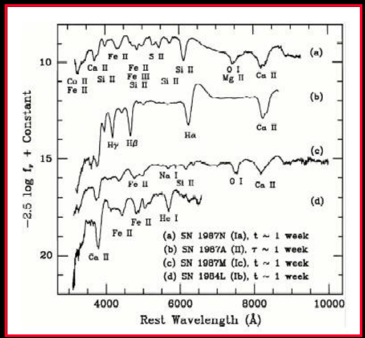


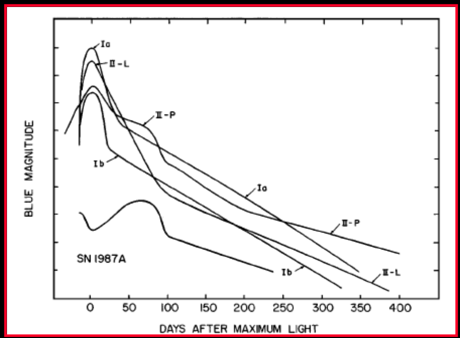
Large Digital Sky Surveys

- Ripe problems for machine learning in astrophysics
- Growing source of data in astronomy: ~ 10 TB of image data, ~ 10^9 detected sources, ~ 10^2 measured attributes per source
- Increasingly heterogeneous data

Examples

- Search for transients: supernovae, GRB afterglows, fainter and faster phenomena
- SNe classification: standardizing spectra and light curves
- Weak Gravitational Lensing: measuring galaxy shapes





Contributions

- Demonstrated potential of machine learning to have high scientific impact
  - Integration into nightly operations for supernova search
  - Prototype for future digital sky surveys
- Advantage of classifiers
  - Can model various data sets, e.g. different telescopes, different lunar phases, searches for other transient objects
  - Adapt to equipment calibrations, image processing software modifications by retraining
  - Handle large, imbalanced data sets

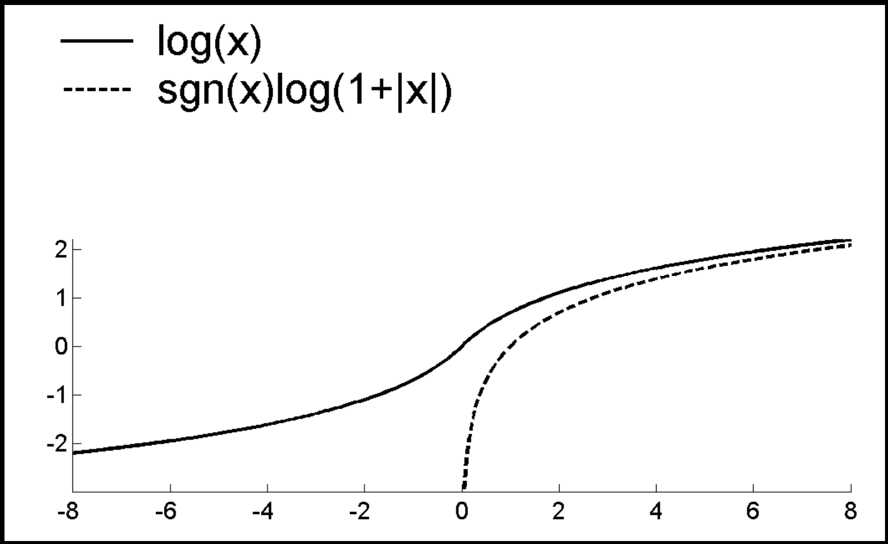
END

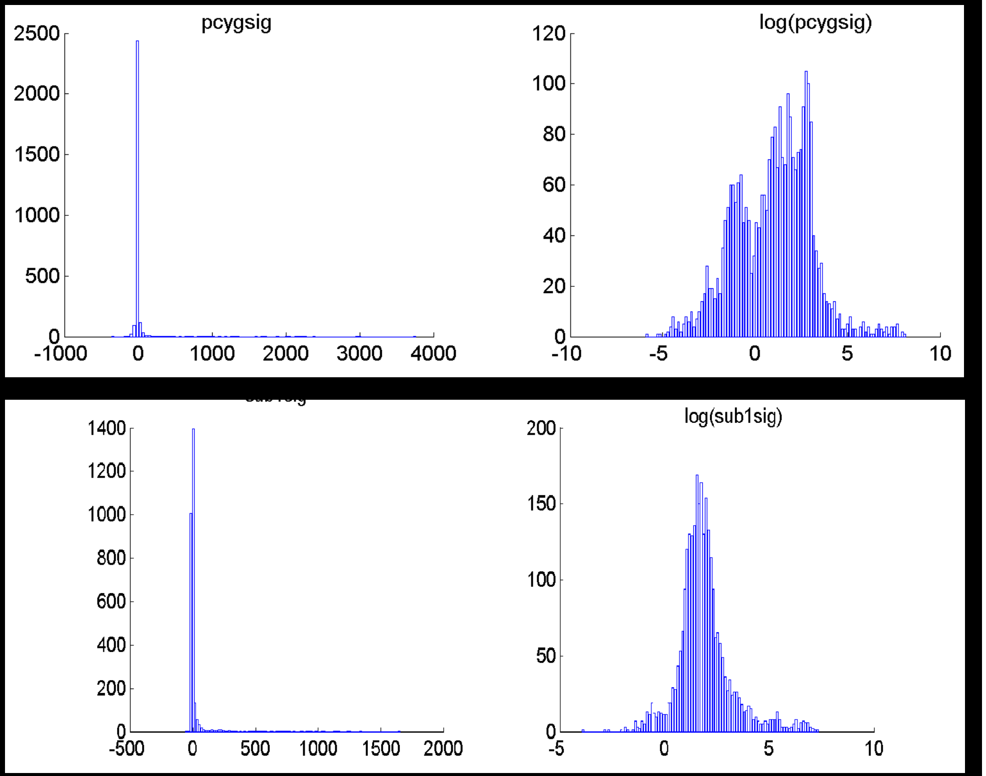
Joint work with Cecilia Aragon, Chris Ding, and The Nearby Supernova Factory at LBNL

Supernova Recognition using Support Vector Machines, R. Romano, C. Aragon, and C. Ding, International Conference of Machine Learning Applications, December 14-16, 2006. To appear.

Feature Transformation

- Highly peaked, skewed distributions
- May take on negative values
- Transforming some dimensions may change path of optimization





SNFactory Features

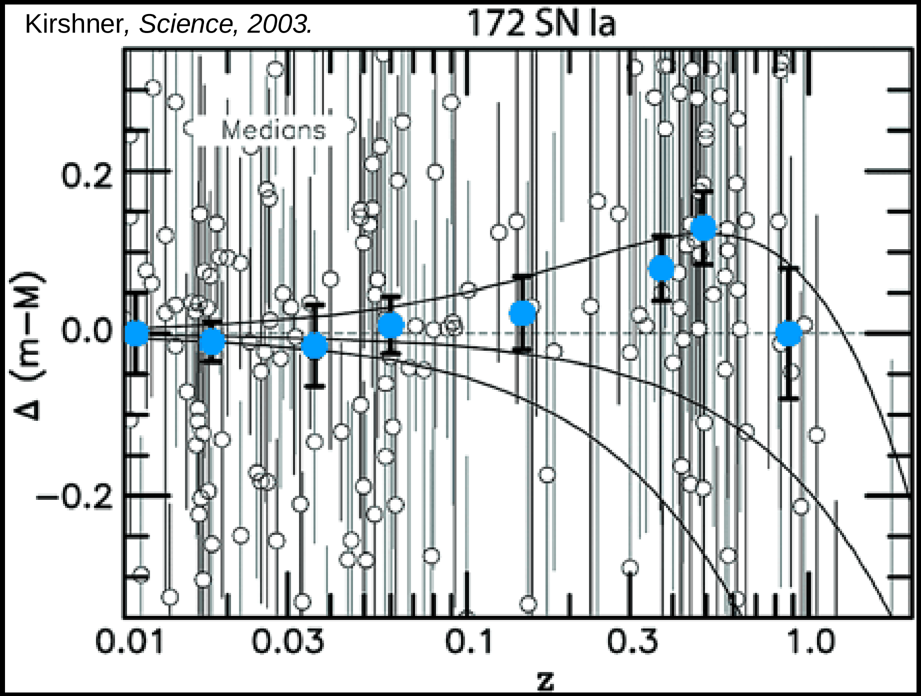
Feature Name	Feature Definition
apsig	signal-to-noise ratio in aperture
perinc	% flux increase in aperture from REF to NEW
pcygsig	difference of flux in 2*FWHM of aperture and 0.7*FWHM; detects misaligned REF and NEW images)
mxy	x-y moment of candidate
fwx	FWHM of candidate in x
fwy	FWHM of candidate in y
neighbordist	distance to the nearest object in REF
new1sig	signal-to-noise of candidate in NEW1
new2sig	signal-to-noise of candidate in NEW2
sub1sig	signal-to-noise of candidate in SUB1
sub2sig	signal-to-noise of candidate in SUB2
sub2minsub1	weighted signal-to-noise difference between SUB1 and SUB2
dsub1sub2	difference in pixel coordinates between SUB1 and SUB2 (motion measurement)
holeinref	measure of negative pixels on REF in region of candidate
bigapratio	ratio of sum of positive pixels to sum of negative pixels within aperture
relfwx	REF image FWHM in x divided by NEW image FWHM in x
relfwy	REF image FWHM in y divided by NEW image FWHM in y roundness object contour eccentricity; ratio of powers in lowest order negative and positive Fourier contour descriptors
wiggliness	object contour irregularity; power in higher order Fourier contour descriptors divided by total power

Type Ia Supernovae Studies

- Stellar explosions appearing as bright spots near galaxies
- Uniform peak brightness, spectra, and temporal light curves
- Spectroscopic measurements provide direct experimental evidence that universe is accelerating
- Time-varying spectra of *thousands* of Type Ia supernovae needed to constrain estimate of acceleration rate
- Difficulty: rare (1-2 per millenium), random, and fleeting (several weeks)

galaxy

galaxy with supernova



Lagrangian Dual

- Begin with the primal

$$f(x) \text{ s.t. } g(x) \leq 0$$

- Take annoying constraints into objective function with multipliers

$$L(u) = f(x) + u^T g(x)$$

- Minimize L for the given u

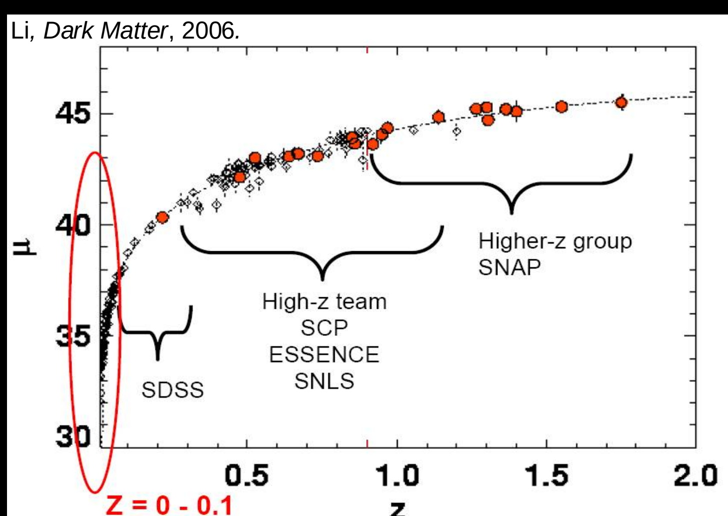
$$L^*(u) = \min_x f(x) + u^T g(x)$$

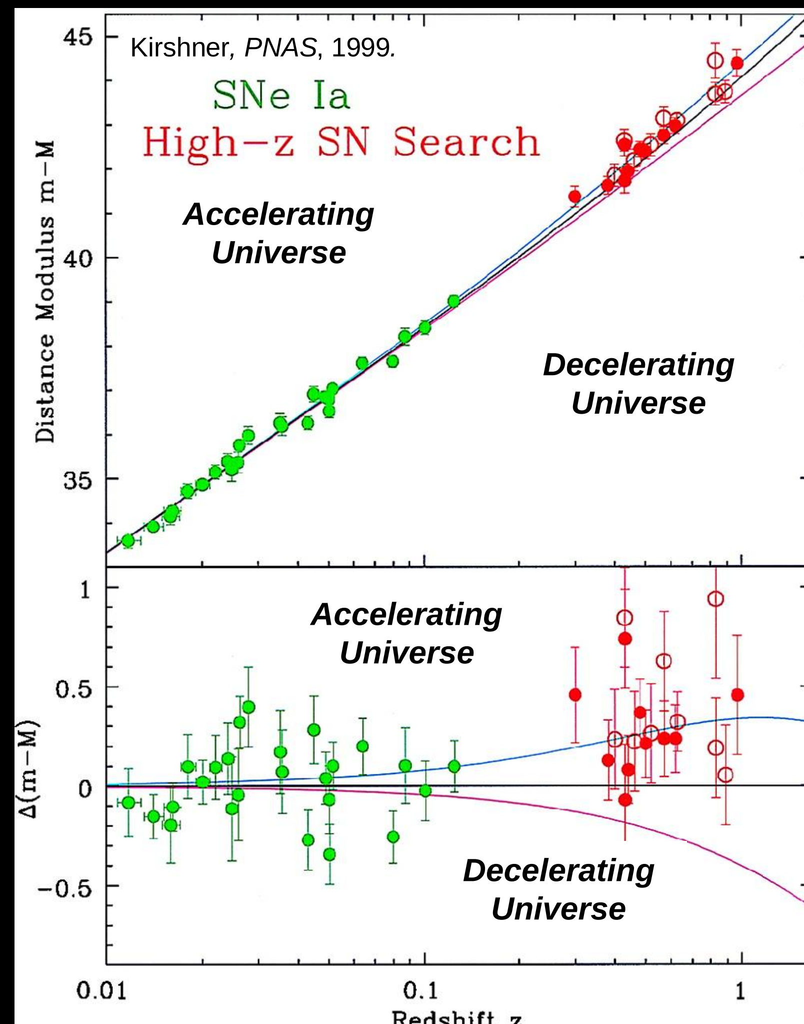
- Assuming that was an easy minimization, maximize over all positive u

$$v^* = \max_{u \geq o} L^*(u)$$

Type Ia Supernovae Studies

- First direct experimental evidence that universe is accelerating
- Uniform peak brightness, optical spectra, and light curves
- Time-varying spectra of thousands of Type Ia supernovae needed to constrain estimate of acceleration rate



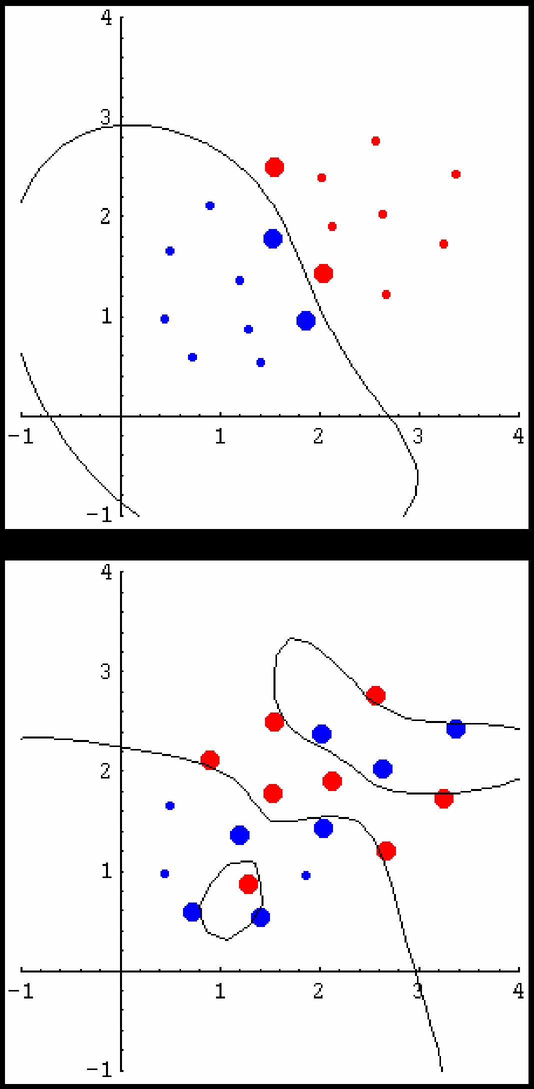


Details to Worry About

- Overfitting:

  trade-off between complexity of boundary and generalization error

- Parameter-selection:
  - variance on Gaussian kernels
  - constants on soft margin terms in objective function: how much slack to allow



More Astrophysics Applications: Spectra Classification